**Gene-Centered Analysis of Sets of Genes and Data Integration Applied to the Discovery of Presumptive Cis-Regulatory Sites in the Genome**

***Kirov, Stefan*[*1,6], *Zhang, Bing*[1,6], *Crawford, Oakley*[1,6], *Peng, Xinxia*[1,6], *Williams, Robert*[2], *Langston, Michael*[1,3], *Leuze, Michael*[1,4], *Jones, Brynn*[1,4], *Larimer, Frank*[1,6], *Baker, Erich*[7], *Goldowitz, Dan*[2], *Snoddy, Jay*[1,6], *Tennessee Mouse Genome Consortium, Integrative Neuroscience Initiative on Alcoholism, and The Program for Comparative and Collaborative Bioinformatics***

***[1]University of Tennessee and Oak Ridge National Laboratory Graduate School for Genome Science and Technology; [2]The University of Tennessee Health Science Center, Memphis, TN, USA; [3]The University of Tennessee, Computer Science Department, Knoxville, TN, USA; [4]Oak Ridge National Laboratory, Computer Science and Mathematics Division, Oak Ridge, TN, USA; [5]Oak Ridge National Laboratory, Functional Genomics Group, Oak Ridge, TN, USA; [6]Oak Ridge National Laboratory, Genome Analysis and Systems Modeling Group, Oak Ridge, TN, USA; [7]Baylor University, USA***

**Biology Requirements and Drivers:** We need new ways for computers to process the sets of genes and the data sets that can be associated with those genes, especially the sets of genes that are networked together in important developmental and physiological processes. In particular, we need to compare sets of genes and associated data across different species, across different genotypes, across different cell types, and across different physiological states. There is a need to have genome-scale computational analysis much more automated as is usually possible rather than "one gene and one click at the time" web interface approach. We need new data mining environments and computational pipelines that can scale up to work with large data sets, let alone the intersections and unions across several large gene and data sets. One area of particular interest is to understand how a set of genes whose transcription may be co-regulated, presumably via shared and overlapping sets of cis-regulatory sites.

**Overall Approach and Results to Date:** The bioinformation systems we are building include several major features (see also abstracts of Snoddy et al and Zhang et al.). This poster will describe several approaches to genome analysis, with a special emphasis to combining information from different data sets to try to deduce presumptive cis-regulatory sites in co-regulated genes. We have developed a system, called **GeneKeyDB.** This is a gene-centered database that can automatically integrate different data about overlapping sets of genes and data that can be associated with genes or gene products. GeneKeyDB is implemented as a lightweight, gene-centric relational database which combines data from various existing resources, such as LocusLink, CGAP, HomoloGene, Gene Ontology, and ENSEMBL. GeneKeyDB consists of keys (IDs), and pointers to the methods to get the archival data behind the keys, rather than trying to store the actual data inside this database. The system is lightweight enough to be copied and used in various analysis tools on different computers. **BSA** (Batch Sequence Analysis) is just one set of systems that can use GeneKeyDB as a data integrator and data mining environment on a single computer or in a more distributed computing grid. The BSA pipeline can analyze sets of data that we believe from experimental data or from hypothesis to be co-regulated, and the pipeline can try to find a small set (4-10) of potential cis-regulatory sites that may confer co-regulation in genes that QTL or other analysis might suggest are co-regulated (e.g. work of the WebQTL project). GeneKeyDB assists BSA to also see if those cis-regulatory elements are conserved in other chordate genomes. We will present current results from GeneKeyDB and BSA at this meeting that can strongly suggest several putative cis-regulatory sites in several different gene sets.